CRESST REPORT 831

AUTOMATIC SHORT ESSAY SCORING USING NATURAL LANGUAGE PROCESSING TO EXTRACT SEMANTIC INFORMATION IN THE FORM OF PROPOSITIONS

AUGUST, 2013

Deirdre Kerr

Hamid Mousavi

Markus R. Iseli

National Center for Research
on Evaluation, Standards, & Student Testing

UCLA | Graduate School of Education & Information Studies

Automatic Short Essay Scoring Using Natural Language Processing to Extract Semantic Information in the Form of Propositions

CRESST Report 831

Deirdre Kerr
CRESST/University of California, Los Angeles

Hamid Mousavi
CSD/University of California, Los Angeles

Markus R. Iseli
CRESST/University of California, Los Angeles

August 2013

National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
Center for the Study of Evaluation (CSE)
Graduate School of Education & Information Studies
University of California, Los Angeles
300 Charles E. Young Drive North
GSE&IS Bldg., Box 951522
Los Angeles, CA 90095-1522
(310) 206-1532

The work reported herein was supported by grant number OPP1003019 from The Bill and Melinda Gates Foundation with funding to the National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

The findings and opinions expressed in this report are those of the authors and do not necessarily reflect the positions or policies of The Bill and Melinda Gates Foundation.

To cite from this report, please use the following as your APA reference: Kerr, D., Mousavi, H., & Iseli, M. R. (2013). *Automatic short essay scoring using natural language processing to extract semantic information in the form of propositions* (CRESST Report 831). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

TABLE OF CONTENTS

AUTOMATIC SHORT ESSAY SCORING USING NATURAL LANGUAGE PROCESSING TO EXTRACT SEMANTIC INFORMATION IN THE FORM OF PROPOSITIONS

Deirdre Kerr
CRESST/University of California, Los Angeles

Hamid Mousavi
CSD/University of California, Los Angeles

Markus R. Iseli
CRESST/University of California, Los Angeles

Abstract

The Common Core assessments emphasize short essay constructed-response items over multiple-choice items because they are more precise measures of understanding. However, such items are too costly and time consuming to be used in national assessments unless a way to score them automatically can be found. Current automatic essay-scoring techniques are inappropriate for scoring the content of an essay because they either rely on grammatical measures of quality or machine learning techniques, neither of which identify statements of meaning (propositions) in the text. In this report, we introduce a novel technique for using domain-independent, deep natural language processing techniques to automatically extract meaning from student essays in the form of propositions and match the extracted propositions to the expected response. The empirical results indicate that our technique is able to accurately extract propositions from student short essays, reaching moderate agreement with human rater scores.

Introduction

The impending implementation of Common Core assessments across the United States brings with it a shift in large-scale assessments from multiple-choice items to short essay constructed-response items as measures of student knowledge of a given concept (Wu, 2012). Short essay constructed-response items consist of a targeted prompt, such as "Explain how fractions and decimals are related," and the written response of each student to that prompt. These items measure deeper understanding (Baker, Aschbacher, Niemi, & Sato, 1992) and are more precise than multiple-choice items in measuring a student's understanding of a given topic (Huang, Tsai, Hsu, & Pan, 2006; Jacobs-Lawson & Hershey, 2002; Klein, Chung, Osmundson, & Herl, 2002). However, short essay items are also significantly more difficult to score than multiple-choice items because the need for human raters makes the process both time consuming and expensive (Magliano & Graesser, 2012; Wu, 2012).

Automated, unsupervised methods of scoring student textual responses would significantly reduce the heavy workload currently associated with large-scale scoring of constructed-response items (Villalon & Calvo, 2009), as well as the often prohibitive costs of such scoring (O'Neil & Klein, 1997; Rozali, Hassan, & Zamin, 2011). Additionally, automatically extracting content could mitigate the subjectivity of human raters and reduce bias against poorly written works that are conceptually correct (Smith & Humphries, 2006), which could result in more accurate assessments of knowledge for students who are English Language Learners or who struggle with writing.

However, automatically scoring the accuracy of students' textual responses is a particularly challenging problem (Graesser, McNamara, & Louwerse, 2010; Valerio & Leake, 2006). Research to date has focused largely on either grading essays based on grammar, coherence, and style or grading the content of short-answer questions based on bag-of-words approaches or machine learning techniques (Shermis, Burstein, Higgins, & Zechner, 2010; Landauer, Laham, & Foltz, 2003; Mohler & Mihalcea, 2009).

These approaches are increasingly successful at identifying the quality of writing and the general content area being covered in the text, but they cannot extract the basic statements of meaning, or *propositions* (Kintsch, 1974), that indicate the precise level of conceptual understanding evinced in a given essay. If student textual responses are going to be used as large-scale assessments of student content knowledge in areas such as math or science, where writing quality is not the metric of concern, the automatic extraction of propositions from student texts is necessary.

In this report, we explore the possibility of using a deep Natural Language Processing (NLP) based technique to extract propositions from student text and utilize graph matching techniques to compare extracted propositions to propositions from a target essay in order to automatically score the semantic content of short-answer constructed-response items.

Related Work

An alternative methodology to scoring essays is based on scoring essay-derived knowledge maps by comparing them to expert-created knowledge maps. Research has shown that using knowledge maps to assess conceptual understanding can be a valid and robust alternative approach to essay scoring (Chung et al., 2003; O'Neil & Chung, 2011; Ruiz-Primo, Shavelson, Li, & Schultz, 2001). Advances in natural language processing by our group and others have made the extraction of propositions (i.e., <*concept, relation, concept*> tuples) tractable (Lajis & Aziz, 2010; Mousavi, Kerr, & Iseli, 2011; Pérez-Marín, Alfonseca, Rodríguez, & Pascual-Nieto, 2007; Valerio & Leake, 2006).

One of the most common methods of extracting propositions from text is a lightweight summarization approach (Vargas-Vera & Moreale, 2005) that uses part-of-speech tagging to identify nouns in the text and then uses machine learning techniques to determine whether or not there is a semantic relationship between the identified nouns based on the relative proximity of the nouns to each other and the frequency of their co-occurrence in the same sentence. Cañas et al. (2005), Chen, Kinshuk, Wei, and Chen (2008), Gaines and Shaw (1994), Lau, Song, Li, Cheung, and Hao (2009), Smith and Humphreys (2006), and Tseng, Chang, Rundgren, and Rundgren (2010) all use this approach. However, this lightweight summarization approach often results in unlabeled relationships (Huang et al., 2006), because the proximity and co-occurrence measures used in these lightweight summarization approaches allow for the identification of the existence of a link between nouns but do not provide information about the link. Some methods expand on this technique by adding information from dependency graphs to access more linguistic information, but this process is still not fully linguistic (Bailey & Meurers, 2008; Mohler, Bunescu, & Mihalcea, 2011).

Fully linguistic methods of proposition extraction are more accurate than these statistical lightweight summarization methods, but are also less common due to the difficulty involved in implementing them efficiently and effectively (Vargas-Vera & Moreale, 2005). However, linguistically based studies often place artificial constraints on the processes to achieve an acceptable level of accuracy in link identification. These constraints result in the identification of only a subset of the propositions stated in the text. Some studies constrain the links to a predetermined set of verbal relationships (Valerio & Leake, 2006) or use ontological domain information to constrain the verbal relationships for each noun (Richardson, Srinivasan, & Fox, 2008), while others constrain the verbal relationships to those in which nouns in the content area of interest are the subject of the link (Zouaq & Nkambou, 2008).

Constraints that limit the verbal relationships to a predefined set of relationships or to information stored in a specific ontology result in techniques that are domain-specific, as the constraints applied in one domain often do not generalize to other domains (Kowata, Cury, & Boeres, 2010). The few linguistic studies that were domain-independent applied linguistic methods only at the concept extraction stage, while the link extraction stage remained based upon co-occurrence (Tseng et al., 2010; Wang, Cheung, Lee, & Knok, 2008).

We propose a fully linguistic method of extracting propositions from text. Though the process uses an ontology to help determine which nodes and links are within the targeted content area, the ontology generation process is automated so the process remains domain-independent and does not require human intervention.

Data Set

The corpus used in this study consisted of a preexisting, pre-scored set of short essay responses by fourth- and fifth-grade students explaining the hearing process. The students replied to the following prompt:

> Imagine your friend comes to you with a problem. She has missed the last two months of school and wants you to explain how ears work. You need to explain all about the ear and the hearing process.
>
> Think about all of the important things you've learned about hearing and how our ears work. Also think about the relationships between the different parts of the ear and how the ear as a whole goes together. Then write an explanation to your friend so that she can understand hearing.

In total, the corpus consisted of 55 short essays containing approximately 5100 words in 415 sentences. Each essay was scored on a 1 to 5 scale by two raters using a holistic rubric, where a score of 5 indicated an essay that covered all the main scientific principles on the rubric and contained no conceptual errors and a score of 1 indicated an essay that covered none of the main scientific principles. Interrater reliability was high ($\alpha = .95$), and where disagreements occurred a consensus was reached on a final score, rather than taking the mean of the two different scores (Klein et al., 2002). Only one student received a score of 1, but the remaining scores were fairly evenly distributed, with a mean score of 3.53 and a standard deviation of 1.02.

The scientific principles listed in the rubric were in the form of complete sentences. These sentences were combined in paragraph form to create the *target essay* for the automatic scoring process:

> The outer ear (auricle) catches sound waves. Sound waves travel through the ear canal to vibrate the eardrum. The vibrating eardrum passes vibrations on to the middle ear, which is made up of the hammer, anvil, and stirrup. The middle ear passes vibrations to the inner ear, via the stirrup. The stirrup vibrates the oval window. The vibrating oval window causes the fluids in the cochlea to vibrate. The cochlea converts the sound waves to electrical impulses via the fibers in the cochlea. The electrical impulses from the cochlea travel to the brain via the auditory nerve. The brain interprets the vibration as sound.

Proposition Extraction in SemScape

The SemScape framework (Mousavi et al., 2011) extracts propositions from free text using grammatical information present in the text. The grammatical information is first identified using the Stanford parser (Stanford NLP Group, 2013) which converts each sentence into a parse tree that can be indexed so each word's relative position can be easily identified. As shown in Figure 1, once a sentence has been parsed and indexed, SemScape uses a three-step process to mine

propositions from the text: main-part identification, TextGraph generation, and proposition extraction (Mousavi et al., 2011).

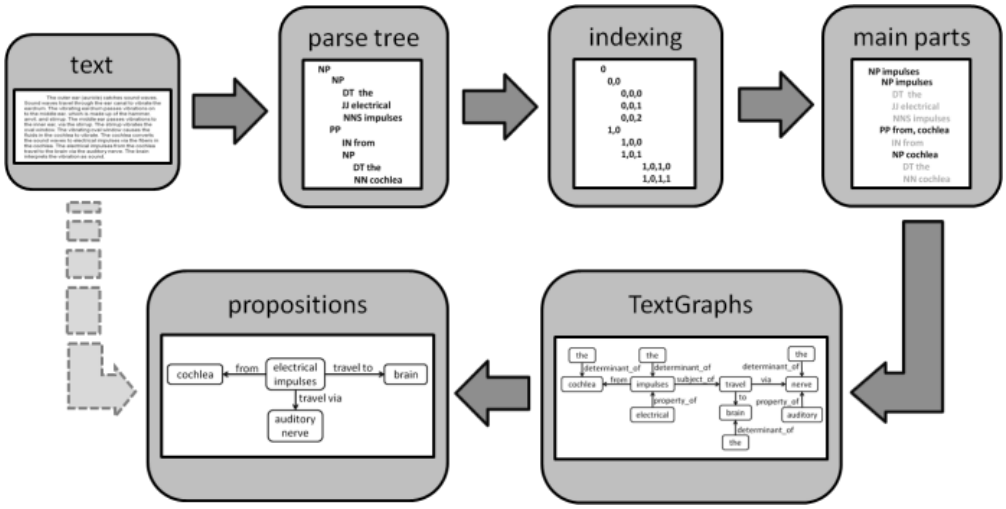

Figure 1. SemScape proposition extraction process.

Given the example sentence, "*The electrical impulses from the cochlea travel to the brain via the auditory nerve*," the indexed parse tree that SemScape would work with is shown in Figure 2.

Index	Parse Tree
-	(S
0	(NP
0, 0	(NP
0, 0, 0	(DT) the
0, 0, 1	(JJ) electrical
0, 0, 2	(NNS) impulses)
0, 1	(PP
0, 1, 0	(IN) from
0, 1, 1	(NP
0, 1, 1, 0	(DT) the
0, 1, 1, 1	(NN) cochlea)))
1	(VP
1, 0	(VBP) travel
1, 1	(PP
1, 1, 0	(TO) to
1, 1, 1	(NP
1, 1, 1, 0	(NP
1, 1, 1, 0, 0	(DT) the
1, 1, 1, 0, 1	(NN) brain)
1, 1, 1, 1	(PP
1, 1, 1, 1, 0	(IN) via
1, 1, 1, 1, 1	(NP
1, 1, 1, 1, 1, 0	(DT) the
1, 1, 1, 1, 1, 1	(JJ) auditory
1, 1, 1, 1, 1, 2	(NN) nerve))))))

Figure 2. Example of an indexed parse tree.

Main-Part Identification

The main-part of each phrase (noun phrase, verb phrase, prepositional phrase, etc.) is identified using approximately 130 main-part rules that examine the grammatical structure of the parse tree and use that information to carry the main-parts of the leaves of the parse tree up to the parent nodes. For example, the rule in Figure 3 examines branches in a noun phrase (NP). If the branch includes a determinant (DT), followed by an adjective or adjective phrase (JJ|ADJP), followed by a noun (NN|NNS), the noun located at index 0,2 will be copied up to the noun phrase in index 0.

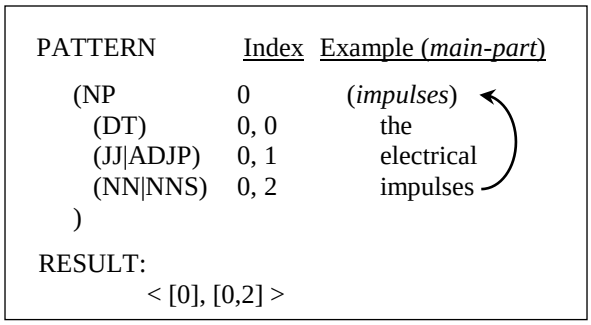

Figure 3. An example of a main-part rule.

This rule would identify the main-part of the noun phrase "*the electrical impulses*" as "*impulses.*" Carrying the main-part information up to the branches allows subsequent rule sets to be far more parsimonious because variations in leaf structures for each branch are already accounted for. Additionally, given access to an ontology (see the Ontology Integration section of this report) main-part rules can identify multi-word terms as well. For example, the hearing ontology would allow the main-parts of the noun phrase "*the electrical impulses*" to be identified as both "*electrical impulses*" and "*impulses.*"

TextGraph Generation

The grammatical relationships between words in the text are identified using approximately 270 tree domain rules that take advantage of parse tree and main-parts information. These relationships are converted into a *TextGraph* wherein the nodes are words in the sentence and the links are the grammatical relationships between those words. For example, the rule shown in Figure 4 would identify every occurrence of a noun phrase (NP) followed directly by a verb phrase (VP) and link the main-part of the noun phrase to the main-part of the verb phrase with the link "*subject_of.*"

PATTERN	Index	Example (*main-part*)
(NP)	0	(*electrical impulses*)
(VP)	1	(*travel*)

RESULT:
< [0], subjectOf, [1] >

Figure 4. An example of a text domain rule.

This is a lossless process meant to extract the grammatical structure of the sentence exactly as it is written. Semantic meaning is not inferred at this point and, because the rules are entirely based on grammar, they do not have to be modified for different conceptual domains, though they would have to be modified for styles of writing that do not follow the same grammatical structures, such as poetry or dialogues. Given the example sentence "*The electrical impulses from the cochlea travel to the brain via the auditory nerve*," the tree domain rules would convert the sentence into the TextGraph shown in Figure 5.

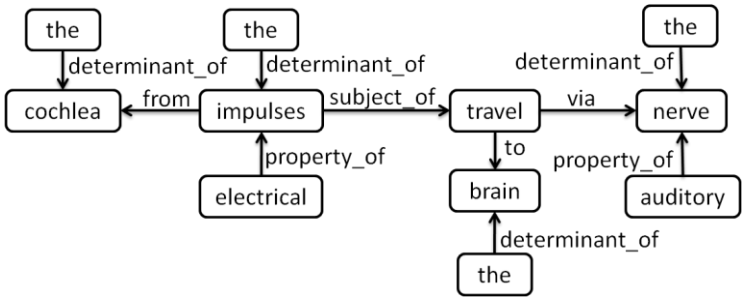

Figure 5. TextGraph for the example sentence.

Ontology Integration

An ontology for the hearing process was generated to provide SemScape with a list of terms in the topic area that would allow for the identification of multi-word terms and the separation of scientific concepts from non-scientific terms.

The ontology was generated automatically using OntoHarvester (Mousavi, Kerr, & Iseli, 2013). OntoHarvester starts with an initial ontology (a *seed*) and iteratively extends the seed using graph-based patterns on TextGraphs. The seed for the hearing ontology consisted of 12 common hearing concepts: *ear*, *eardrum*, *vibration*, *hammer*, *anvil*, *stirrup*, *oval window*, *cochlea*, *brain*, *sound*, *cells*, and *sensory cells*. The corpus fed to OntoHarvester to create the hearing ontology consisted of a chapter about the hearing process from a seventh grade science textbook (Berwald et al., 2007).

OntoHarvester identifies taxonomical relationships (such as *Part_Of* or *Type_Of*) between terms already in the seed and other terms in the corpus. These terms are added to the seed, and another pass is run to identify additional taxonomical relationships between the terms in the expanded seed and other terms in the corpus. The iteration continues until no additional terms are found, at which point the ontology is considered complete.

This process resulted in a final hearing ontology of 81 concepts which was used during proposition extraction to discriminate between conceptual terms like "*electrical impulses*" and descriptions of terms such as "*faint sounds*," and during the matching and scoring process to identify concepts that were directly related to the hearing process but were not specifically mentioned in the target essay.

Proposition Extraction

Propositions are extracted from the TextGraphs using approximately 50 graph domain rules, which, like all our other rules, are entirely based on grammar and therefore are domain-independent. These rules identify patterns in the TextGraph and translate them into <*node, link, node*> triples. For example, the rule in Figure 6 finds every subgraph in the TextGraph wherein a noun (?1) linked to a verb (?3) with the relationship "*subject_of*" and a second noun (?2) is linked to the same verb (?3) with the relationship "*to,*" provided that the word "*not*" does not modify the verb.

```
SELECT (?1 ?3 ?2)
WHERE {
        ?1 "subject_of" ?3.
        ?2 "to" ?3.
                NOT("not" "property_of" ?3).}
```

Figure 6. An example of a graph domain rule.

For the TextGraph in Figure 5, this rule would select <*electrical impulses*> and <*brain*> and extract the proposition <*electrical impulses, travel to, brain*>. All propositions for the example sentence are displayed in graphical form in Figure 7. Note that the above rule also generates the proposition <*impulses, travel to, brain*>, which aids in the process of matching propositions.

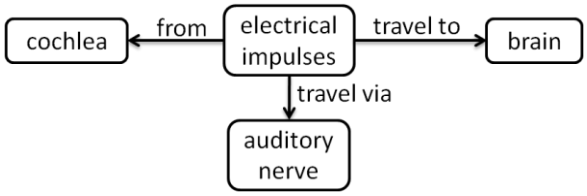

Figure 7. Proposition graph for the example sentence.

Matching and Scoring Process

Once the target essay and the student essays have been converted into proposition graphs consisting of connected <*node*, *link*, *node*> triples in which the *link* specifies the semantic connection between the *nodes*, the proposition graphs are matched to the target essay. Since proposition graphs do not contain information about grammatical structure of the text, the matching of proposition graph triples to triples in the target essay is an easier task than matching either the text or the TextGraphs.

Currently, four features of the extracted triples are considered in the process of aligning the propositions of students' essays to those of the target essay. These features are nodes, concepts, links, and triples. Concepts are defined as the nodes that match terms in the hearing ontology. The process for matching each feature is as follows:

Nodes: Before starting the matching, all the nodes are converted to their singular forms. For each node, say *n*, in a student essay, we say the node matches the target essay if:

- There is an exact matching node in the target essay.
- One of the synonyms of *n* listed in the hearing ontology is used in the target essay (e.g., "*labyrinth"* in a student essay matches "*cochlea"* in the target essay).
- One of the hypernyms of *n* listed in WordNet (Princeton University, 2010) is found in the target essay (e.g., "*tube"* in student essay matches "*cochlea"* in the target essay).

Concepts: For each node in a student essay, we say there is a concept match if the matching node in the target essay is a concept in the ontology.

Links: Similar to nodes, we first convert verbs to their infinitive form in order to simplify the matching process. For each link *l* in the student essay, we say *l* matches the target essay if:

- There is an exactly matching link in the target essay.
- The main verb in *l* is a troponym or synonym of the main verb of a link (*l'*) in the target essay, while all other parts of *l* and *l'* match each other. For instance "*go to"* in student essay matches "*travel to"* in the target essay, since "*go*" is a troponym for "*travel*" in WordNet (Princeton University, 2010).

Triples: We say a triple <n_1, *l*, n_2> in a student essay matches the target essay if there exists a triple <m_1, *k*, m_2> such that i) nodes n_1 and n_2 respectively match nodes m_1 and m_2, and ii) link *l* matches link *k*. This allows exact matching as well as near matching such as <*wave*, *goes in*, *middle ear*> and <*sound waves*, *travel in*, *middle ear*>.

Combining the Evidence

Two measures of each of the features used in the scoring process were taken into consideration in the final score of each essay (see Table 1). A *count* feature indicated the number of occurrences of the feature in each essay and a *match* feature indicated the number of matches to the target essay.

Table 1

Features Used to Compute Essay Scores

Feature	Description
Node count	Number of unique nodes in the essay
Concept count	Number of unique concepts in the essay
Link count	Number of unique links in the essay
Triple count	Number of unique triples in the essay
Node match	Number of nodes in the essay that match nodes in the target
Concept match	Number of concepts in the essay that match concepts in the target
Link match	Number of links in the essay that match links in the target
Triple match	The number of triples in the essay that match triples in the target

Obviously, match features are better indicators of the conceptual accuracy of the extracted propositions than count features. Therefore, match features were used as the primary indicators of the final score of each essay except for the special case of an essay with a concept count of 0. Essays with a concept count of 0 were given a final score of 1 because they did not discuss the topic of interest. For example, the lone essay scoring a 1 in this data set discussed how to solve a dispute on the playground rather than describing the hearing process.

To determine which of the four remaining scores on the scale (2, 3, 4, or 5) to assign to each essay with a concept count greater than 0, the percentage of essays receiving each score from the human raters was calculated. In this data set, 18% of essays received a score of 2, 20% of essays received a score of 3, 45% of essays received a score of 4, and 15% of essays received a score of 5. All four match feature scores were broken into ranks corresponding to the percentage of essays falling into each score category. Then the mean of the match feature ranks was calculated and rounded to the nearest whole number to get a mean match score.

This mean match score was then based on the length and breadth of the essay, as measured by the four count features. These features were also broken into ranks corresponding to the percentage of essays falling into each score category. Then the difference between the mean

match feature and the ranking of each count feature was calculated. If the sum of those differences was more than two, the essay's score was increased by one point to give credit for the additional explanation. If all four differences were negative, the essays' score was decreased by one point for being off topic. The score alignment performance of each feature score, the mean match score, and the final essay score adjusted for length can be seen in Table 2.

Table 2

Score Alignment Performance of Each Feature

Feature score	Correlation	RMSE	Kappa
Node match rank	.686	.813	.398
Concept match rank	.595	.932	.333
Link match rank	.680	.813	.323
Triple match rank	.569	.952	.298
Node count rank	.647	.869	.282
Concept count rank	.621	.901	.402
Link count rank	.631	.890	.317
Triple count rank	.551	.991	.262
Mean match score	.678	.824	.384
Final essay score	.719	.777	.473

Results

To evaluate SemScape's ability to extract semantic information from student responses to short essay prompts, we tested two components of the scoring technique: proposition extraction and final score alignment. To measure the performance of the proposition extraction process, we report precision and recall for the extracted propositions. To measure the performance of the final score alignment process, we report the Pearson's correlation coefficient and Root Mean Square Error (RMSE) as recommended in Ziai, Ott, and Meurers (2012), as well as the interrater reliability coefficient Cohen's Kappa (a common measure of interrater reliability that takes into account the probability of scoring each item correctly by chance).

Proposition Extraction Performance

To evaluate the first phase of our algorithm, we manually verified the correctness of the generated propositions for the first 11 essays (20%) in our data set. The summary of the results is shown in Table 3. As can be seen in the Precision column, the average accuracy of the generated propositions is more than 76%. This is a very impressive result considering that, as is common

with students of this age, there are a number of grammatical mistakes in the essays. Many of these mistakes result in the generation of incorrect parse trees, which in turn affects our results.

To compute the recall of the generated propositions (column three of Table 3), we compared the number of automatically extracted propositions to the manually generated set of propositions for each essay. Recall values range between 52% and 88%, with an average of 63%.

Table 3

Hand Scoring Results for 20% of Essays

Essay	Human score	Recall	Precision	Error
1	3	69%	84%	16%
2	5	53%	80%	20%
3	5	48%	66%	34%
4	4	64%	80%	20%
5	4	67%	82%	18%
6	4	54%	48%	52%
7	4	70%	85%	15%
8	3	57%	83%	16%
9	5	63%	68%	32%
10	4	52%	92%	8%
11	3	88%	62%	38%
Overall	-	63%	76%	24%

Approximately 15% of the error observed in our proposition extraction is due to incorrect Anaphora/Pronoun Resolution. This is a particular problem with student writing because incorrect pronoun usage interferes with most methods of pronoun resolution. For example, the pronoun in the sentences "*The sound waves enter the ear. Then it goes down the ear canal.*" is particularly difficult to resolve because the singular pronoun "*it*" is meant to refer to the plural noun "*sound waves*" which is farther away from the pronoun than the singular noun "*ear.*" This greatly increases the chance of resolving the pronoun to the wrong term.

It is important to note that the proposition extraction error does not appear to be correlated with the score the essay received from the human rater. This means that we are not systematically mischaracterizing either low-scoring or high scoring essays. However, since the essay score was not based on writing quality, it is possible that the error rate is systematically mischaracterizing students based on writing style or quality.

Final Score Alignment Performance

The final scores generated by SemScape were compared to the scores assigned by human raters. The Pearson correlation between the computer rater scores and the scores given by the human raters was .719 and the RMSE was .777 (see Table 2). These values are better (higher correlation, lower RMSE) than those reported in other similar studies such as Mohler et al. (2011) and Ziai et al. (2012). However, the Cohen's Kappa measure of interrater reliability was only .473, where values below .4 indicate a weak agreement, values between .4 and .6 indicate moderate agreement, values between .6 and .8 indicate substantial agreement, and values above .8 indicate the raters are interchangeable.

Table 4

Comparison of Computer Rater to Human Raters

Human score	Computer score					Percent matched
	1	2	3	4	5	
1	**1**	-	-	-	-	100%
2	-	**6**	1	3	-	60%
3	-	3	**4**	4	-	36%
4	-	1	3	**18**	3	72%
5	-	-	-	2	**6**	75%

Table 4 shows the distribution of computer rater scores to human rater scores. There were only three scores where the computer was off by more than one point, scoring three essays as 4's rather than 2's and one essay as a 2 instead of a 4. In all other cases, the computer rater was within one point of the human rater. However, the percentage of matched scores indicates that the computer rater had difficulty accurately scoring essays that received 3's from human raters, correctly scoring only 36% of those scores while performing much better for the other scores.

Conclusions and Future Work

This report demonstrates a technique for using domain-independent, deep natural language processing to score short essay responses by automatically extracting propositions from student writing using SemScape. This technique successfully extracted propositions from student essays, achieving an average recall of 63% and an average precision of 76%. It also successfully replicated the scores given by human raters, achieving a correlation of .719, an RMSE of .777, and a Cohen's Kappa of .473. Future work will focus on increasing the precision and recall of the proposition extraction process by adding additional graph domain rules to identify semantic

relationships contained in less common grammatical structures. More importantly, the scoring process will be significantly enhanced by adding subgraphs as a feature used for matching. This addition would allow for the identification of a match that is written in one sentence in the target essay but is split across two sentences in the student essay. Additionally, more careful weighting of the individual features might provide a better match and help raise the Cohen's Kappa above .8 so that the computer can be considered a reliable rater of short answer constructed responses.

References

Bailey, S., & Meurers, D. (2008). Diagnosing meaning errors in short answers to reading comprehension questions. In *Proceedings of the Third ACL Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 107-115). Columbus, OH: Association for Computational Linguistics.

Baker, E. L., Aschbacher, P. R., Niemi, D., & Sato, E. (1992). *CRESST performance assessment models: Assessing content area explanations* (CSE Report 652). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Berwald, J., et al. (2007). *Focus on life science, California, Grade 7* (Student Edition). Columbus, OH: Glencoe/McGraw-Hill.

Cañas, A. J., Carff, R., Hill, G., Carvalho, M., Arguedas, M., Eskridge, T. C., … Carvajal, R. (2005). Concept maps: Integrating knowledge and information visualization. *Lecture Notes in Computer Science*, *3426*, 181-184.

Chen, N.-S., Kinshuk, Wei, C.-W., & Chen, H.-J. (2008). Mining e-learning domain concept map from academic articles. *Computers & Education*, *50*(3), 1009-1021.

Chung, G. K. W. K., Baker, E. L., Brill, D. G., Sinha, R., Saadat, F., & Bewley, W. L. (2003). Automated assessment of domain knowledge with online knowledge mapping. *Proceedings of the I/ITSEC*, *25*, 1168–1179.

Gaines, B. R., & Shaw, M. L. G. (1994). Using knowledge acquisition and representation tools to support scientific communities. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI '94)* (pp.707-714). Menlo Park, CA: American Association for Artificial Intelligence.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2010). Methods of automated text analysis. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of Reading Research* (pp. 34-53). New York, NY: Routledge.

Huang, C.-J., Tsai, P.-H., Hsu, C.-L., & Pan, R.-C. (2006). Exploring cognitive differences in instructional outcomes using text mining technology. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC 2006)* (pp. 2116-2120). Taipai, Taiwan: IEEE Computer Society.

Jacobs-Lawson, J. M., & Hershey, D. A. (2002). Concept maps as an assessment tool in psychology courses. *Teaching of Psychology*, *29*, 25-29.

Kintsch, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

Klein, D. C. D., Chung, G. K. W. K., Osmundson, E., & Herl, H. E. (2002). *Examining the validity of knowledge mapping as a measure of elementary students' scientific understanding* (CSE Report 557). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Kowata, J. H., Cury, D., & Boeres, M. C. S. (2010). Concept maps core elements candidates recognition from text. In J. Sánchez, A. J. Cañas, & J. D. Novak (Eds.), *Proceedings of the Fourth International Conference on Concept Mapping* (pp. 120-127). Santiago, Chile: Universidad de Chile.

Lajis, A., & Aziz, N.A. (2010). NL scoring technique for the assessment of learners' understanding. *Proceedings of the 2nd International Conference on Computer Research and Development* (pp. 379–383). Washington, DC: IEEE.

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Erlbaum.

Lau, R. Y. K., Song, D., Li, Y., Cheung, T. C. H., & Hao, J.-X. (2009). Toward a fuzzy domain ontology extraction method for adaptive e-learning. *IEEE Transactions on Knowledge and Data Engineering, 21*(6), 800-813.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed response. *Behavioral Research, 44*, 608-621.

Mohler, M., & Mihalcea, R. (2009). Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2009)* (pp. 567-575). Stroudsburg, PA: Association for Computational Linguistics.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp.752-762). Stroudsburg, PA: Association for Computational Linguistics.

Mousavi, H., Kerr, D., & Iseli, M. R. (2011). *A new framework for textual information mining over parse trees* (CRESST Report 805). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Mousavi, H., Kerr, D., & Iseli, M. R. (2013). *Unsupervised ontology generation from unstructured text* (CRESST Report 827). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

O'Neil, H. F., & Chung, G. K. W. K. (2011, April). *Use of knowledge mapping in computer-based assessment.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

O'Neil, H. F. Jr., & Klein, D. C. D. (1997). *Feasibility of machine scoring of concept maps* (CSE Technical Report 460). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Pérez-Marín, E., Alfonseca, E., Rodríguez, P., & Pascual-Nieto, I. (2007). Automatic generation of students' conceptual models from answers in plain text. In C. Conati, K. McCoy, & G. Paliouras (Eds.), *User Modeling 2007* (pp. 329-333). Berlin: Springer-Verlag.

Princeton University. (2010). WordNet [Computer Software]. Retrieved from http://wordnet.princeton.edu

Richardson, W. R., Srinivasan, V., & Fox, E. A. (2008). Knowledge discovery in digital libraries of electronic theses and dissertations: An NDLTD case study. *International Journal on Digital Libraries, 9*, 163-171.

Rozali, D. S., Hassan, M. F., & Zamin, N. (2011). Development of preprocessing modules for an adaptive qualitative assessment using with dynamic question generation: A concept map approach. In A. Patil & C. S. Nair (Eds.), *Proceedings of the International Engineering and Technology Education Conference (IETEC '11)*. Kuala Lumpur, Malaysia: IETEC.

Ruiz-Primo, M. A., Shavelson, R. J., Li, M., & Schultz, S. E. (2001). On the validity of cognitive interpretations of scores from alternative concept-mapping techniques. *Educational Assessment, 7,* 99–141.

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed.). Oxford, UK: Elsevier.

Smith, A. E. & Humphreys, M. S. (2006). Evaluation of unsupervised semantic mapping of natural languages with Leximancer concept mapping. *Behavior Research Methods, 38*(2), 262-279.

Stanford NLP Group. (2013). The Stanford parser: A statistical parser [Computer Software]. Retrieved from http://nlp.standford.edu/software/lex-parser.shtml

Tseng, Y.-H., Chang, C.-Y., Rundgren, S.-N. C., & Rundgren, C.-J. (2010). Mining concept maps from news stories for measuring civic scientific literacy in media. *Computers & Education, 55*(1), 165-177.

Valerio, A., & Leake, D. (2006). Jump-starting concept map construction with knowledge extracted from documents. In A. J. Cañas & J. D. Novak (Eds.), *Proceedings of the Second International Conference on Concept Mapping (CMC '06)* (pp. 296-303). San José, Costa Rica: Universidad de Costa Rica.

Vargas-Vera, M., & Moreale, E. (2005). Automatic extraction of knowledge from student essays. *International Journal of Knowledge and Learning, 1*(4), 318-331.

Villalon, J. J., & Calvo, R. A. (2009). Concept extraction from student essays, toward concept map mining. In I. Aedo, N.-S. Chen, Kinshuk, D. Sampson, & L. Zaitseva (Eds.), *Proceedings of the Ninth IEEE International Conference on Advanced Learning Technologies (ICALT 2009)* (pp. 221-225). Washington, DC: IEEE Computer Society.

Wang, W. M., Cheung, C. F., Lee, W. B., & Knok, S. K. (2008). Mining knowledge from natural language texts using fuzzy associated concept mapping. *Information Processing & Management, 44*(5), 1707-1719.

Wu, H.-S. (2012). Assessment for the Common Core mathematics standards. *Journal of Mathematics Education at Teachers College, 3*, 6-18.

Ziai, R., Ott, N., & Meurers, D. (2012). Short answer assessment: Establishing links between research strands. In *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications 2012* (pp. 190-200). Red Hook, NY: Curran Associates.

Zouaq, A., & Nkambou, R. (2008). Building domain ontologies from text for educational purposes. *IEEE Transactions on Learning Technologies, 1*(1), 49-62.